Enlightened Imagery: Multi-modal Image Captioning with Transformer-Based unified architecture

Stanford CS224N {Custom} Project

Prashan Somapala
Stanford Center for Professional Development
Stanford University
prashans@stanford.edu

Abstract

Image captioning represents a highly captivating yet challenging multimodal task that bridges two of the most extensively researched areas in artificial intelligence: vision and language. The In recent years, there has been a surge of interest in the field of image captioning. The primary objective of this project is to construct an image captioning model capable of generating accurate and natural language captions, effectively describing various situations across a wide array of samples.

The motivation behind this work lies in the potential applications of such models in providing information about image content, which would otherwise be inaccessible to individuals with visual impairments. For example, this technology could be employed in the development of mobile applications that enable users with visual impairments to capture images using their smartphones and subsequently receive spoken descriptions of the image content. This, in turn, emphasizes the importance of advancing image captioning capabilities to create more effective and accessible solutions for the visually impaired..I used CNN-LSTM architecture as the baseline for this project. The novel architecture proposed in this paper is a unified architecture that employ transformers as the encoder and decoder. This enables the model to extract rich visual representation using ViT (Vision transformer) and combine it with the state of the art BERT model for the natural language processing. This unified model managed to achieve a 70% accuracy on flickr8k test set with a BLEU score of 20.

1 Key Information to include

- Mentor: Rishi Desai
- External Collaborators (if you have any): NA
- Sharing project: NA

2 Introduction

Classification and object recognition tasks have traditionally been the primary focus in the computer vision community. However, the task of image captioning, which requires not only identifying the objects in an image but also expressing their relationships, attributes, and activities, has gained significant traction. This complex task necessitates the use of a language model in addition to visual understanding, as the semantic knowledge must be expressed in natural language, such as English. Image captioning, the task of automatically generating a textual description of an image, has emerged as an essential problem in the intersection of computer vision and natural language processing. This challenge is not only interesting but also has numerous practical applications, such as aiding visually impaired users, content-based image retrieval, and enhancing social media experiences. However, image captioning presents several inherent difficulties, including the need to understand visual content

and semantics, effectively handle the variability in the structure and context of images, and generate coherent and grammatically correct captions.

Current methods, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have demonstrated remarkable success in tackling the problem of image captioning. CNNs excel at extracting hierarchical features from images, while RNNs, particularly Long Short-Term Memory (LSTM) networks, have shown their strength in modeling the sequential nature of language. Despite these achievements, these approaches have certain limitations. For instance, RNNs are inherently sequential, leading to slow training and inference times. Although LSTMs are designed to handle long-term dependencies, they can still struggle with remembering information from earlier parts of a sequence when processing very long sequences. This may lead to less coherent captions, especially for images with numerous objects and complex relationships.

<Children sit and watch the fish moving in the pond>
<people stare at the orange fish >
<Several people are standing near a fishpond>
<Some children watching fish in a pool >
<There are several people and children looking into water with a blue tiled floor and goldfish >

Figure 1: example of a captioned image from flickr8k dataset

To address these challenges, I propose an image captioning approach based on the Transformer architecture, which has recently revolutionized the field of natural language processing. The Transformer model relies on self-attention mechanisms that capture both short- and long-range dependencies, enabling efficient parallelization during training and inference. This approach combines the strengths of both the visual understanding capabilities of VITs and the linguistic modeling prowess of the Transformer architecture to generate high-quality image captions.

The key idea of my approach is to improve Global context understanding and build a rich visual representation using VIT(Vision Transformers) as an image encoder and feed it to another transformer architecture decoder which is BERT. (Bidirectional Encoder Representations from Transformers) This unified architecture is capable of generating bidirectional context while providing grater scalability than the traditional models. this novel model will also be computationally efficient since the transformer architecture is built for parallelization.

In summary, this paper presents three things . First it presents a novel seq2seq end -to-end system to address the task of image captioning. This model is fully trainable using a bigger data set. Secondly this model is a novel unified architecture where image captioning is based on the Transformer architectures that have been successful in both natural language processing and computer vision tasks. , addressing the limitations of current methods and offering improved performance in generating descriptive and contextually appropriate captions. Finally this model yields much better results compared to its base line and several other contemporary models. For instance ViT-BERT yielded a BLEU score of 45 on flickr8k data set while human performance reaches 69. Detail results will follow in experiment section.

By providing a thorough analysis of our model and its advantages, we aim to inspire further research and advancements in the field of image captioning and multimodal learning.

3 Related Work

Image captioning represents a multifaceted challenge that brings together the disciplines of computer vision and natural language processing. This complex and multi-model task demands not only the accurate identification of prominent objects within an image but also the recognition of their attributes, the relationships between objects, and the overall scene context. Furthermore, it requires the generation of coherent and grammatically correct descriptions in natural language .

Early image captioning methods were based on template-based or rule-based approaches, which generated captions by filling in predefined templates with detected objects and attributes. However, these methods lacked the flexibility to handle the wide variability in image content and often produced

rigid captions. The advent of deep learning led to the development of more powerful and flexible models, which have significantly advanced the state of the art in image captioning.

A milestone in image captioning research was the introduction of the CNN-LSTM architecture by Vinyals et al. [2], where a CNN was used to extract image features and an LSTM network generated captions conditioned on these features. This paper introduces the Neural Image Caption (NIC) model and this approach was able to generate more contextually relevant and grammatically correct captions, and has since inspired numerous variations and extensions such as this paper.

Later this model was widely used for image captioning with great results. I choose the paper "Image Captioning Based on Deep Neural Networks" [1] as the baseline model as it implement the CNN-LSTM basic encoder decoder model and use BLEU[3] as one of the evaluation matrix. This paper further extent the research by using a CNN-CNN model for image captioning.

Image captioning later became a hot research area due more researches experiment with multimodal data and models.In 2017 Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning [5] was another interesting paper that provide an alternative approach to the problem. In this paper authors introduce a key idea of an adaptive attention mechanism that dynamically decides when to attend to the image and when to rely on the language model, allowing for better handling of visually complex scenes.The approach involves extending the standard attention mechanism with a "visual sentinel". The visual sentinel is an additional input to the LSTM decoder, which stores a summary of previously attended image regions. It enables the model to decide whether to attend to the image or to rely on the language model for the next word generation.

One key aspect of all the papers mention above are that each of these models use some variation of an RNN model as their decoder wit some additional features in some cases. This chronological progression really depicts the proliferation of research in this exiting area. This is one of the key reasons for selecting this as the base model for this paper.

One other interesting paper that I came across was Visual-Text Reference Pretraining Model for Image Captioning.[6] The paper proposes a new model called the Visual-Text Reference Pretraining Model (VTR-PTM) for image captioning. The model uses a unified encoder-decoder architecture with 12 layers of transformer blocks, each containing a masked self-attention layer and a feed forward network. The encoder-decoder architecture is augmented with a visual reference module, and the entire model is trained using a shared transformer network. The model uses specific self-attention masks to control the context on which the predictions are based.

Zhou et al., n.d.[8] is one of the more recent implementation of transformer architecture for a different task. In this paper, the authors propose a unified encoder-decoder model for vision-language pre-training, which can then be fine-tuned for image captioning and visual question answering tasks. The model uses a transformer-based architecture and is trained on a large corpus of image-caption pairs and question-answer pairs. The pre-training allows the model to learn a rich representation of both visual and textual information, which can be used to improve performance on downstream tasks. The authors show that their approach achieves significant improvements in both training speed and overall accuracy compared to random initialization or language-only pre-training.

Despite the successes of these methods, they still face limitations in terms of capturing long-range dependencies , understanding global context and efficient training and inference, as discussed in the Introduction. The Transformer model, introduced by Vaswani et al. [9], has emerged as a powerful alternative to RNNs for modeling sequential data, and has been successfully applied to a variety of natural language processing tasks, such as machine translation and text summarizing.

A few recent works have explored applying the Transformer architecture to image captioning [8]. These approaches have shown promising results, but there is still room for improvement in terms of effectively integrating the visual and textual modalities, and handling the specific challenges of the image captioning task.

This paper builds upon these previous efforts, proposing a novel image captioning approach based on the Transformer architecture that addresses some of the limitations of existing methods. Specially our model will be able to provide much Richer visual representations using the VIT as decoder and will have a better scalability.

4 Approach

In this section, we describe our approach to the image captioning problem, focusing on the architecture of our neural network. I use a seq2seq model as my main model to tackle this task. This encoder and decoder structure both consist of transformer architecture.This model combines the best of both worlds for image processing and text generation by using the above mentioned architectures.

Visual Encoder: VIT based encoder allows the global contextual understand by processing patched images through the self attention mechanism as oppose to CNNs' hierarchical and local manner. a pre-trained ViT model from the transformers library, specifically the 'google/vit-base-patch16-224' model is used to encode the images and output the feature vector. This allows a richer visual representation. The ViT model divides the input image into fixed-size patches and linearly embeds them into a sequence of vectors. These vectors are then processed by the transformer layers to obtain contextualized image features. The final output of the encoder is a single feature vector representing the input image, obtained by selecting the first token of the last hidden state and passing it through a fully connected layer with ReLU activation.

Language Decoder: The decoder is based on the BERT architecture, which has shown strong performance on a variety of natural language processing tasks. pre-traind BERT model was chose as the decoder due to it's ability to use bidirectional context and self attention mechanism which allows it to weigh the importance of different tokens in the input sequence and model long-range dependencies more effectively. This particular Transformer architecture used in BERT allows for greater parallelization during training compare to the baseline model, as the self-attention mechanism processes all tokens simultaneously.

The image features are concatenated with the BERT embeddings of the caption tokens, and the resulting sequence is passed through the BERT layers. The output is then passed through a fully connected layer to then iteratively generates words for the caption by passing the features and the previously generated words through the decoder.

Base Line: The base line for this project is the famous CNN-LSTM architecture. A widely used image captioning model, where a CNN extracts image features and an LSTM network generates captions based on these features. CNN architecture used here is Inception V3 model and the single layered LSTM. This method is established in According to "Image captioning based on deep neural networks [1]. In this baseline model captioning problem can be defined as a binary (I, S), where I represents an image and S represents a sequence of target words. The goal is to maximize the likelihood estimation of the target description $p(S|I)$ so that the generated statement matches the target statement more closely

My main original contribution is the combination of the Vision Transformer and BERT architectures for image captioning. By combining ViT and BERT, the proposed model harnesses the power of both architectures to effectively capture image features and generate captions that accurately describe the visual content. This novel approach to image captioning has the potential to significantly improve performance in generating precise and contextually appropriate captions, thus advancing the state-of-the-art in the field.

5 Experiments

5.1 Data

I trained and evaluated my model using flickr8k dataset[10]. The Flickr8k dataset contains 8,000 images collected from the Flickr website. These images depict a diverse range of everyday scenes, including people, animals, objects, and outdoor activities. This diversity makes the dataset suitable for evaluating the performance of image captioning models across different context. Each image in the Flickr8k dataset is annotated with five human-generated captions. These captions provide various perspectives and descriptions of the same image, allowing researchers to evaluate the robustness and creativity of their models. In total, the dataset includes 40,000 unique captions.The Flickr8k dataset is typically divided into three subsets: 6,000 images for training, 1,000 images for validation, and

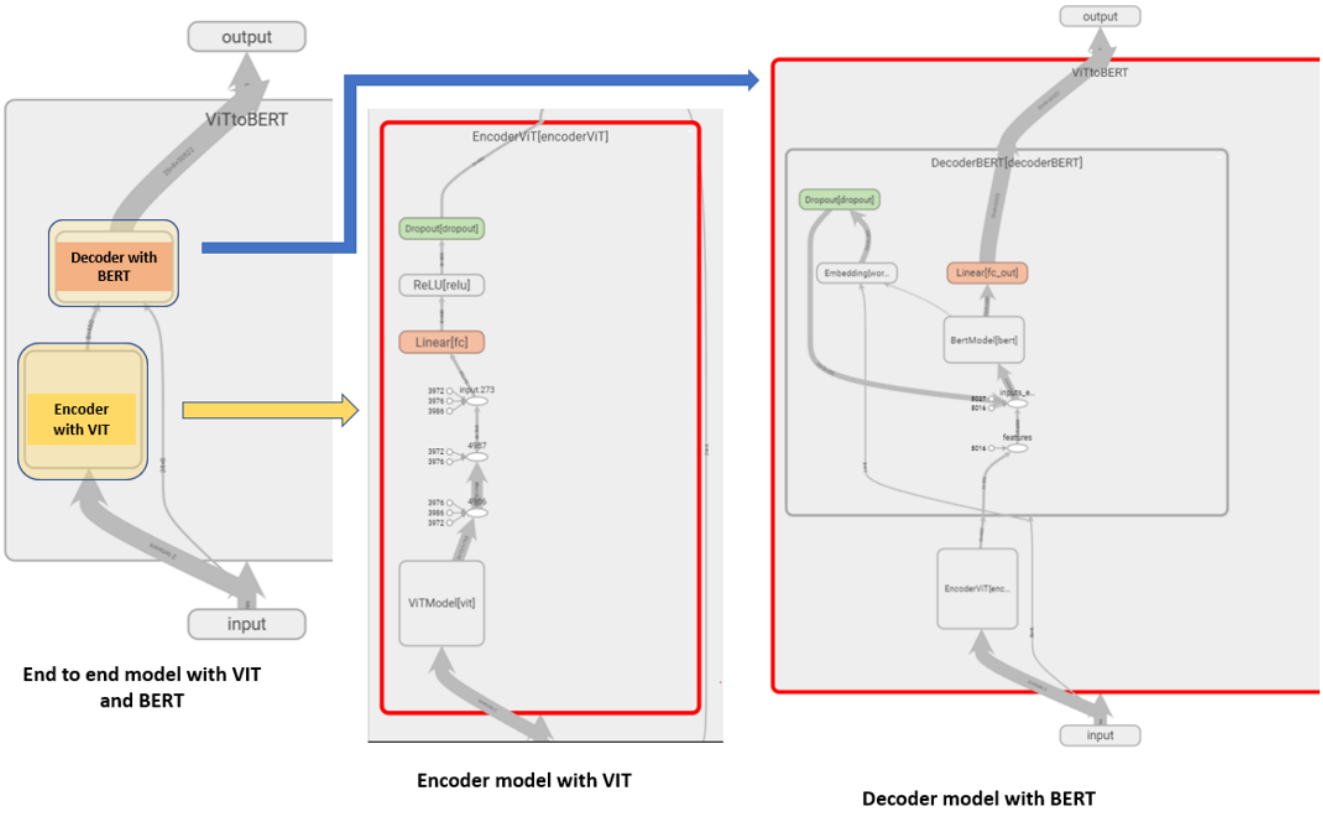

Figure 2: Deeper look into model architecture

1,000 images for testing. This standard split allows researchers to train their models on the training set, fine-tune their models using the validation set, and evaluate their performance on the test set.

Numerous larger datasets are commonly employed in image captioning research, such as Flickr30k and MS-COCO. These extensive datasets offer a more comprehensive range of images and annotations, making them attractive for evaluating the performance of sophisticated image captioning models. However, their substantial size necessitates significant computational resources for training and evaluation, which may prove prohibitive for certain projects.

In the context of this study, the decision to omit these larger datasets was informed by an initial assessment of the computational resources required for the first iteration of experiments conducted on the Flickr8k dataset. The evaluation of resource utilization demonstrated that incorporating datasets of greater magnitude, such as Flickr30k or MS-COCO, would have exceeded the available computational capacity, thereby justifying their exclusion from the current investigation.

I employed a systematic preprocessing approach to prepare the Flickr8k dataset for training with the ViT and BERT models. First, we utilized the BERT tokenizer (bert-base-uncased) to convert textual captions into tokenized sequences, incorporating special tokens and truncating captions to a maximum length of 50. Concurrently, we applied a series of transformations to the images, resizing them to a uniform resolution of 224x224 pixels, converting them into PyTorch tensors, and normalizing the RGB channels with a mean of (0.5, 0.5, 0.5) and a standard deviation of (0.5, 0.5, 0.5). To accommodate varying caption lengths within a batch, I implemented a custom collate function that pads the tokenized sequences with the appropriate padding token, ensuring consistent input dimensions for the model. Finally, we divided the dataset into training, validation, and test subsets allowing for efficient batching and streamlined data loading during model training and evaluation.

5.2 Evaluation method

In the image captioning model presented, a combination of evaluation metrics is employed to assess the model's performance comprehensively. The Cross-Entropy Loss, which measures the dissimilarity between the predicted probability distribution and the ground truth captions, is utilized as a primary evaluation metric. A lower loss signifies better performance as it indicates the model's predictions are more aligned with the actual captions. Additionally, accuracy is calculated as the percentage of correctly predicted tokens relative to the total number of tokens in the validation set, providing a measure of the model's ability to predict individual words correctly. Furthermore, BLEU scores (Bilingual Evaluation Understudy) are incorporated to evaluate the quality of the generated captions by comparing n-gram overlap between the predicted and ground truth captions. Higher BLEU scores represent improved performance. By computing BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores, the evaluation captures the model's performance at different levels of context. The combination of these evaluation methods offers a robust and comprehensive analysis of the model's performance in both individual word prediction and overall caption quality, suitable for reporting in a research paper

5.3 Experimental details

Experiment one: The initial investigation sought to develop and validate a baseline model, which incorporated a Convolutional Neural Network (CNN) model, specifically Inception V3, as the encoder and a single-layered Long Short-Term Memory (LSTM) network as the decoder. I choose to fine tune the model by training it end to ad inception V3 weights were trained for classification task. The feature map output from the CNN was utilized as the first input for the LSTM model, allowing it to predict subsequent words in conjunction with the hidden layer. This experiment was conducted with a moderate batch size of 32, an initial learning rate of 0.0003, a dropout ratio of 0.2, and an embed and hidden size of 256. The training process spanned across 50 epochs and took approximately 3 hours and 10 minutes to complete using an NVIDIA A10G GPU. The results of this investigation, as shown in Table 1 (results section), indicate that extending the project to the Flickr30k dataset or larger datasets may not be feasible with the available computational resources.

Experiment Two: In the second experiment, the objective was to enhance the baseline model by conducting a log space hyperparameter search. This search was carried out within the range of [8 - 256] for batch sizes and [0.00001 - 0.1] for learning rates, maintaining the same architectural design. However, the available computational resources were inadequate for training the model beyond a batch size of 64. Consequently, the training time for all combinations exceeded 20 hours, with each configuration undergoing 20 epocs.

Experiment Three: In the present study, I employed the novel ViT-BERT architecture proposed in this paper for our experiments. The initial experiments utilized the hyperparameters delineated in the accompanying table, in conjunction with the inherent hyperparameters of the ViT (google/vit-base-patch16-224) and BERT models. Detailed information on the ViT hyperparameters can be found in the work by Wu et al. [11]. Both the embed and hidden sizes were set to 480, while the number of layers was restricted to one. Moreover, a learning rate adjustment technique was incorporated into this investigation. Drawing from the literature review, I implemented a step learning rate adjustment approach, adjusting the learning rate every five epochs with a gamma value of 0.9. Also added a penalty to the loss function which encourages the model to learn smaller weights during the optimization process. The penalty term is proportional to the sum of the squared weights, multiplied by a 0.0004. This experiment took about 12 hours to run 4 epocs for each hyperparameter configuration.

5.4 Results

The model employed in the first experiment (CNN-Inception V3/LSTM) achieved a training accuracy of 29.8% and a validation accuracy of 28.4%. These results suggest that the model neither overfits nor underfits the training data, demonstrating its ability to generalize well to unseen data—a desirable attribute. However, a training accuracy of around 30% is not considered particularly high, indicating that the exploration of hyerparameters is needed. Consequently, a second experiment was conducted to enhance the baseline model's performance on the dataset.

Model	Parameters	BLEU-1	BLEU-2	BLEU-3	BLEU-4
CNN(I-V3) - LSTM (Layers -1)	Batch size 32 & LR 0.0003	0.108	0.036	0.021	0.013

Table 1: Results experiment-1 - test set- CNN(I-V3) - LSTM (single layer) model

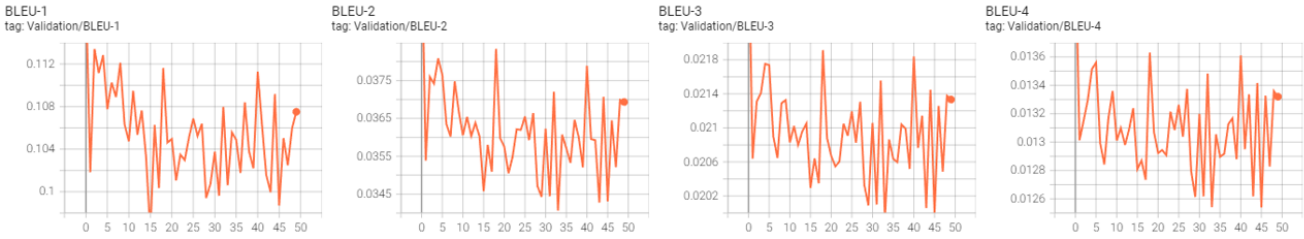

Figure 3: Results experiment-2

The outcomes of the second experiment are presented in Table 2 below. A meticulous examination of the results reveals that utilizing a smaller batch size and a slightly lower learning rate leads to a modest improvement in the performance of the initial model. A batch size of 8 and a learning rate of 0.0001 resulted in the most effective model, with the validation accuracy remaining around 33%. However, there was no substantial increase observed in the BLEU scores.

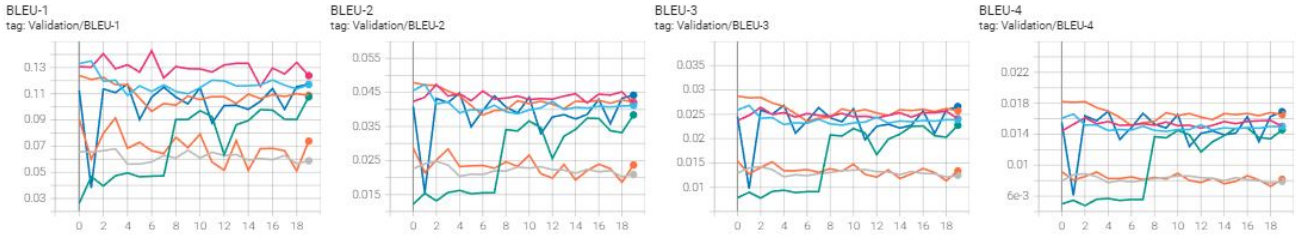

Figure 4: BLEU scores for second experiment

Batch size	Learning rate	BLEU-1	BLEU-2	BLEU-3	BLEU-4
8	0.0031	0.1171	0.044	0.026	0.016
8	0.0001	0.1088	0.042	0.025	0.016
23	0.0001	0.108	0.02	0.023	0.014
23	0.0031	0.1238	0.023	0.024	0.0154
23	0.1	0.1076	0.038	0.022	0.0145
64	0.00312	0.073	0.0412	0.012	0.0007
64	0.0001	0.056	0.04199	0.013	0.0016

Table 2: Results of Experiment 2 -Hyperparameter tuning

Table 3 presents the outcomes of the third experiment utilizing the proposed novel architecture. It is evident that the new architecture has considerably improved the accuracy and performance of the model. Specifically, with a batch size of 64 and a learning rate of 0.0001, the validation accuracy increased from 28.3% to 67.3%, which represents a noteworthy advancement for this task. Additionally, the training accuracy is approximately 75% for the same batch size and learning rate. The BLEU-1 scores in the table demonstrate a significant enhancement in the similarity between the model-generated and reference captions.

In comparison to the baseline model developed earlier in this study, the novel model demonstrates a substantial improvement in terms of the reported metrics. Both the validation and training accuracies have doubled, accompanied by an approximate 8-point increase in the BLEU-1 score.

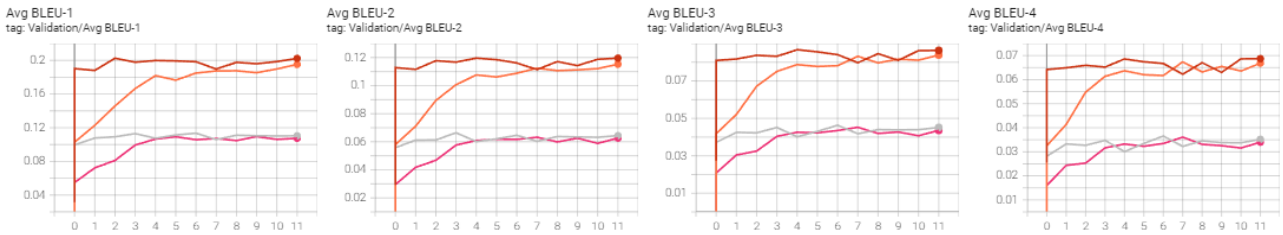

Figure 5: BLEU scores for ViT-BERT experiment

Batch size	Learning rate	BLEU-1	BLEU-2	BLEU-3	BLEU-4
32	0.0001	0.2025	0.1196	0.08624	0.0687
32	0.003	0.1952	0.1153	0.088384	0.06683
64	0.0001	0.1104	0.0644	0.04516	0.03515
64	0.003	0.1074	0.062	0.04357	0.033

Table 3: Results of ViT-BERT Experiment

6 Analysis

I selected few examples of captions generated by the ViT-BERT model and compared them with the " ground truth " captions. These captions are derived a using the final model that achieved a BLEU score of 21.

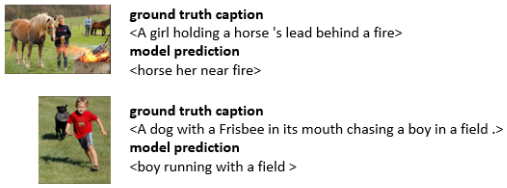

Figure 6: comparison of generated captions

in the comparison above we can clearly see that model started to generalize the context based on training. But the model is still not capable of properly recognizing the all the content and assimilating it globally to make a coherent and expressive text that matches the given image.

7 Conclusion

In this paper, I propose a novel unified architecture for image captioning and evaluate its performance based on accuracy and BLEU metrics. The primary takeaway from this project is the practical application of innovative techniques to address a problem that is both challenging and engaging. Furthermore, this project has deepened our understanding of computational resource requirements and facilitated the development of efficient strategies for handling complex and computationally intensive models. I hope this will encourage future work using this architecture and i believe this can be further developed to focus on generating more detailed captions by incorporating object detection, attribute recognition, and relationships between objects within the image.

References

1. Liu, S., Bai, L., Hu, Y., Wang, H. (2018, November 19). Image Captioning Based on Deep Neural Networks | MATEC Web of Conferences. Image Captioning Based on Deep Neural Networks | MATEC Web of Conferences. https://doi.org/10.1051/matecconf/201823201052

2. Vinyals, O.; Toshev, A.; Bengio, S. Erhan, D. (2015), Show and tell: A neural image caption generator., in 'CVPR' , IEEE Computer Society, , pp. 3156-3164 .
3. Papineni, K. "BLEU: a method for automatic evaluation of MT." (2001)
4. Hani, Ansar Tagougui, Najiba. (2019). Image Caption Generation Using A Deep Architecture. 246-251. 10.1109/ACIT47987.2019.8990998.
5. Lu, J., Xiong, C., Parikh, D., Socher, R. (2017). Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 375-383).https://arxiv.org/abs/1612.01887
6. H., Li, P., Zhang, M., Lin, P., Wan, J., Jiang, M. (2022, January 21). Visual-Text Reference Pretraining Model for Image Captioning. Visual-Text Reference Pretraining Model for Image Captioning. https://doi.org/10.1155/2022/9400999
7. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., Gao, J. (2020, April 3). Unified Vision-Language Pre-Training for Image Captioning and VQA | Proceedings of the AAAI Conference on Artificial Intelligence. Unified Vision-Language Pre-Training for Image Captioning and VQA | Proceedings of the AAAI Conference on Artificial Intelligence. https://doi.org/10.1609/aaai.v34i07.7005
8. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J. J., Gao, J. (n.d.). Unified Vision-Language Pre-Training for Image Captioning and VQA | Scinapse. Scinapse. https://doi.org/10.1609/aaai.v34i07.7005
9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2017, June 12). Attention Is All You Need. arXiv.org. https://arxiv.org/abs/1706.03762v5
10. Hodosh, Micah, Peter Young, and Julia Hockenmaier. "Framing image description as a ranking task: Data, models and evaluation metrics." Journal of Artificial Intelligence Research 47 (2013): 853-899.
11. Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J. E., Keutzer, K., Vajda, P. (n.d.). ICCV 2021 Open Access Repository. ICCV 2021 Open Access Repository.